Telemetry Archive Storage Format for the HST CCS

Paul Ditson

September 25, 1996

Objective

Acknowledge the fact that there are conflicting design requirements for the telemetry archive, some of which are mutually exclusive - compromise is necessary

- Identify the issues
- Characterize the problem
- Discuss several options

Converge on an approach that will strike an acceptable balance

Issues

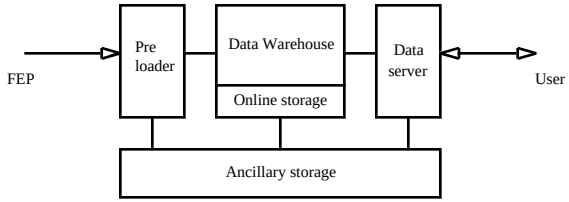
- Volume - cost of:
  - hardware
  - software
  - media
- Performance
  - user response
  - system efficiency
- Complexity
  - preloader and data server
  - warehouse
  - storage management

Requirements

- Volume of data must be manageable and cost effective
- All data points required for spacecraft analysis, mostly near term
- Changes acceptable for near- and long-term trending
- Data should be queryable, retrievals efficient
- Performance must be acceptable for “day 1” queries and for drill down analyses
- Minimize complexity of data server and warehouse support software, maximize COTS
- Encapsulate archive, minimize dependencies in data server and preloader
- All data retrievable for life of mission

System

- Pre-loader - converts captured and merged FOF packets into DW input records
- Data Warehouse - stores, manages, retrieves data, accepts SQL queries from data server
- Data Server - layered between user interface and DW, manages retrieval requests, builds CDF packets, assembles retrievals from distributed sources as necessary



- Ancillary Storage - storage for data which is not directly queryable, but may support retrievals

Options

Six options, each with potential variations:

- All Points
  - Full CDF - 20 bytes
  - Reduced CDF - 13 bytes
- Changes Only
  - All Changes - all mnemonics warehoused
  - Reduced Changes - lo rate w/ averaged hi
- Combined
  - Turbo model - high performance changes, all points archived
  - Economy model - changes archived, all points cached

All Points

- Option 1 - Full CDF
  - 20 byte format - time (8), mnemonic id (2), raw (4), EU (4), flags (2)
  - all information for CDF is contained in warehouse
  - all data points archived for mission
- Option 2 - Reduced CDF (an example)
  - 13 byte reduced format - time code (5), mnemonic id (2), EU (4), flags (2)
  - most information for CDF is contained in warehouse
  - raw analog data stored in ancillary flat files, raw is less likely to be queried
  - objective is to reduce byte count for each data point

Changes Only

- Option 3 - All Changes
  - 20 byte format - time (8), mnemonic id (2), raw (4), EU (4), flags (2)
  - store data point whenever it changes
  - changes include LOS, telemetry format change
  - option 2 can be applied to reduce byte count, with associated trade off
- Option 4 - Reduced Changes
  - 28 byte format - start time (8), stop time (8), mnemonic id (2), raw (4), EU (4), flags (2)
  - store low res data, indexed by change, with time duration
  - selected hi-rate data in flat files
  - selected params averaged, determined by sensor characteristics
  - statistics for all data (min, mean, max, std dev)
  - byte count reduction also an option

Combined - Option 4 + Option 2

- Option 5 - High Performance Model
  - Reduced changes w/ averaged hi rate and statistics kept on RAID, migrated to optical when RAID is saturated
  - All points kept on RAID for short term, migrated to optical for near term online acess, exported to shelf for long term
  - no hi-rate flat files necessary
- Option 6 - Economy Model
  - Reduced changes w/ averaged hi rate and statistics kept in online optical for mission
  - All points on RAID “cache” for near term, no migration to optical
  - hi rate data kept in flat files for mission

Philosophical Divergence

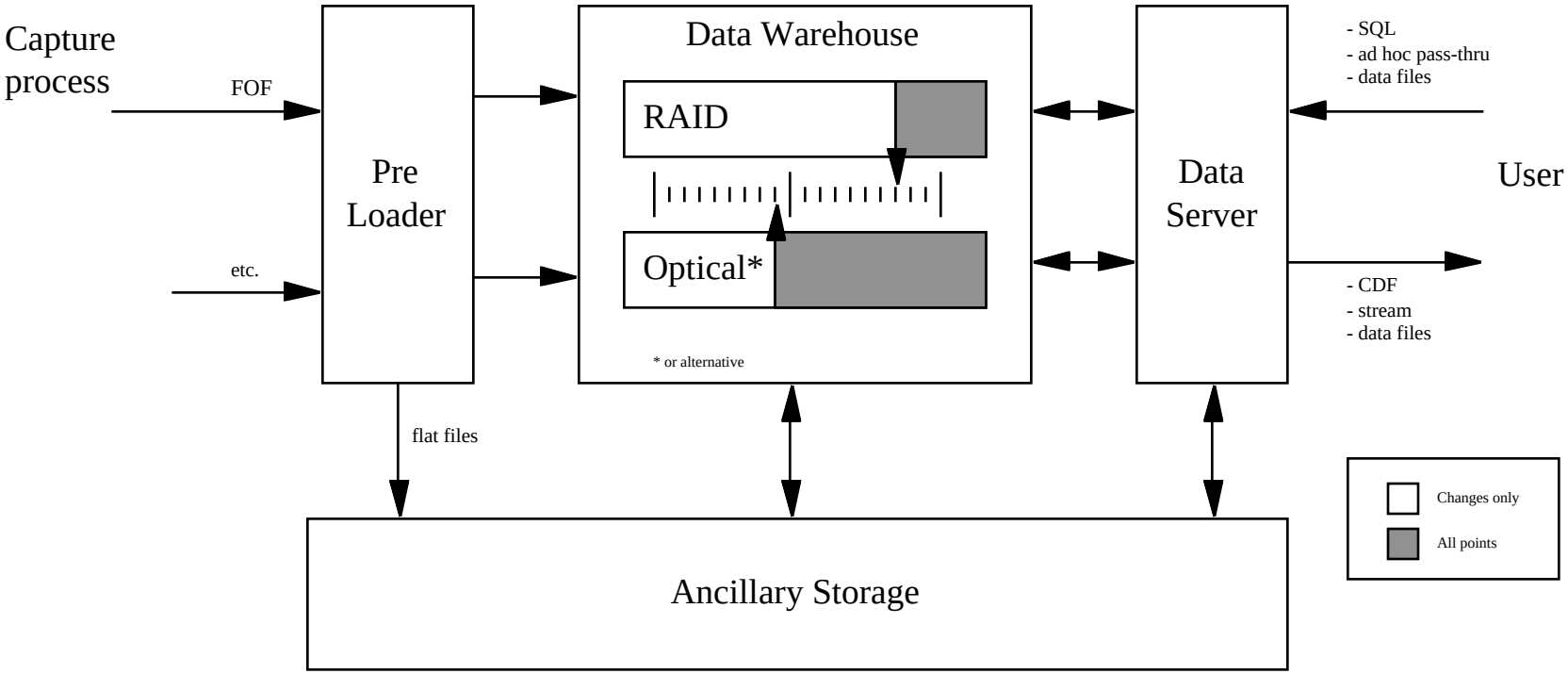
There are two fundamentally opposing uses of the system

- Time domain - what occurred between t_1 and t_2 at various intervals, triggered by event, for several mnemonics
- Frequency domain - how many times has a battery been cycled since launch, has its average depth of discharge changed

- Time domain is interested in meaningful, and flexible, resolution of all points over a time period
- Frequency domain is interested in discrete events or changes, where time is a dependant variable
- Storage and retrieval can be optimized for each, but the solutions are mutually exclusive

A Combined Solution will balance all factors

Conceptual System



All Points - Full CDF

- Volume - 220GB/45TB (month/mission)

  Pro: lowest cost in software development

  Con: highest cost in hardware and media, stores redundant data

- Performance

  Pro: minimal overhead in data server/preloader, direct access to all data, optimized for bounded time-all points

  Con: worst response for Day-1 queries, degrades as time span increases

- Complexity

  Pro: minimal in data server/preloader, low ancillary storage, maximized use of warehouse decision support, direct query capability

  Con: warehouse data mgt highest, incremental backups only

All Points - Reduced CDF

- Volume - 143+16GB/29.2+3.3TB (DW+raw, month/mission)
  - Pro: reduction of 35% DW/27% overall, low-moderate cost of software
  - Con: high cost of h/w and media, redundant data, ancillary storage mgt
- Performance
  - Pro: low overhead in data server/preloader, direct access to all “meaningful” data, optimized for bounded time w/ all points
  - Con: worst response for Day-1 queries, degrades as time span increases, raw data retrieval imparts penalty
- Complexity
  - Pro: low in data server/preloader, maximized use of warehouse decision support, direct query capability
  - Con: ancillary storage increases complexity, DW storage mgt still high

Changes Only - All Changes Warehoused

- Volume - 55GB/11TB (month/mission) (excludes statistics: min, max, mean, std dev)

  Pro: 75% volume reduction (4:1), no ancillary data

  Con: moderate cost of h/w and media, sensitive to spacecraft health/design/signal conditions, no advantage for hi-rate data, additional overhead for periodic snapshots to improve performance

- Performance

  Pro: good for changes-only requests based on time

  Con: poor for solutions requiring all points, where no data exists for start time must “back up” to find most recent change, poor (but not worst) for Day-1 queries

- Complexity

  Pro: less data to process, DB design complexity moderate

  Con: more data in ancillary to manage, lost functionality in warehouse must be implemented in data server (e.g. averages, sampling, sliding windows, hi-res queries), data reconstruction is not trivial

Changes Only - Reduced Changes with Averages

- Volume - 15.3+24GB/3.7+4.9TB (DW+flat, month/mission)

  (excludes statistics; min, max, mean, std dev)

  Pro: lowest volume of data stored in warehouse, reductions of 93% DW/82% overall, lowest cost of hardware/media

  Con: highest DW index overhead due to greater number of tables, highest volume of hi-rate ancillary data, sensitive to spacecraft health/design/data format/signal quality

- Performance

  Pro: optimized for Day-1 and frequency/occurrence based queries, good for averaged queries

  Con: worst performance for time-bounded all-points

- Complexity

  Pro: warehouse data management simplest, less data to process

  Con: DB design most complex, data server must transform changes to all points, complex scheme requires insight into data content, loss of generality, imparts arbitrary rules to storage algorithm, rule changes may require CM and reprocessing of historical data, retrieval from ancillary required for bulk of data, “free” features of DW (stats, avgs, sliding windows, etc) transferred to data server with a cost, assembly of hi-res data from flat files required ...

Combined - Turbo

- Volume - 160+16GB/33+3TB (DW+flat, month/mission)

  Pro: flexible and scalable for any proportion of changes/allpoints, all data permanently archived

  Con: long term management of optical platters, moderate-high volume and hardware/software cost, redundant data stored

- Performance

  Pro: optimized for all queries, capability to cross domains without penalty for complex query, all DW features available

  Con: wrong choice of domain can limit performance, highest preloader loading

- Complexity

  Pro: minimal in data server for most queries, moderate in preloader

  Con: data server must resolve queries which might go to either domain, raw analog reconstruction from flat files

Combined - Economy

- Volume - 1TB cache, 15.3+24GB/3.7+4.9TB (excludes stats, etc.)

  Pro: changes archived for mission, redundant data not stored, low management of optical

  Con: moderate volume and hardware/software cost, requires minor frame archive at FEP to address loss-of-data issue - add approx 2 TB

- Performance

  Pro: optimized for all queries short term, capability to cross domains without penalty for complex query

  Con: wrong choice of domain can limit performance, highest preloader loading, penalty for reconstructing all points historical, limited resolution beyond 180 days, loss of DW features

- Complexity

  Pro: moderate in preloader, minimal in data server for near term - higher beyond cache limit

  Con: data server must resolve queries which might go to either domain, hi-rate queries go to flat files, some complexity shifted to FEP, data server must schedule on-demand ingest

Selection Criteria

	Option 1	2	3	4	5	6
• Complexity 1-lowest, 5-highest						
Preloader						
Data Server						
Warehouse						
Storage Management						
• Performance 1-best, 5-worst						
Ingest						
Retrieve t_1 - t_2						
Retrieve occurrences						
• Cost 1-lowest, 5-highest						
Software						
Hardware						
Media						
Volume DW + flat (GB/month)	220	160	55	40	176	40
(TB/mission)	45	33	11	9	36	9